# Multicodon Nearly Neutral Mutation Model (Stochastic Analysis on Measure-Valued Stocastic Processes)

AUTHOR(S):

Tachida, Hidenori

CITATION:

Tachida, Hidenori. Multicodon Nearly Neutral Mutation Model (Stochastic Analysis on Measure-Valued Stocastic Processes). 数理解析研究所講究録 1999, 1089: 152-165

ISSUE DATE:

1999-03

URL:

http://hdl.handle.net/2433/62851

数理解析研究所講究録
1089 巻 1999 年 152-165



# Multicodon Nearly Neutral Mutation Model

Hidenori Tachida (舘田 英典)
Department of Biology, Kyushu University
email: htachscb@mbox.nc.kyushu-u.ac.jp

February 21, 1999

## Abstract

Evolutionary mechanisms for protein evolution have been intensively studied in the past 30 years. Recent advances in DNA technology provided means to analyze variation at the DNA level and many polymorphism data have been accumulated. These new data provide much more information than protein data and enable us to examine the mechanisms of protein evolution in more detail. Here, I first review what have been learned from these molecular analysis. Then, multicodon nearly neutral mutation models are presented as candidate models for protein evolution and an analysis on one of them is described.

## 1 Introduction

Since KIMURA (1968) proposed the neutral theory of molecular evolution, evolutionary mechanisms at the molecular level have been much debated (see LEWONTIN, 1974; KIMURA, 1983; OHTA, 1996; KREITMAN, 1996). The neutral theory postulates that the main cause of evolutionary change at the molecular level is random fixation of selectively neutral or very nearly neutral mutations rather than Darwinian selection. Especially controversial were the mechanisms for substitutions and maintenance of variation of amino acids. The opposite view to the neutral theory is that of the selectionists in which most amino acid changes are considered to involve adaptive significance (see, for example, GILLESPIE, 1991). As long as just allozyme frequencies were observed, it is difficult to resolve the controversy because the neutral and at least one selection model give the same sampling distribution of gene frequencies in the equilibrium (EWENS, 1972; GILLESPIE, 1977). Also directly measuring differences of fitness within species was very difficult because differences must be very small, say, less than 0.1 % (see MUKAI, ICHINOSE AND TACHIDA, 1981 and DYKHUIZEN AND HARTL, 1980). This was very frustrating for evolutionary biology because proteins are building blocks of organisms and how they evolve and what significance is there among different amino acids have been important questions.

Advances in molecular biology give some hope to break this impasse. In 1970s and 80s, techniques to deal with DNA, the genetic material itself, have been developed and now it is possible to obtain DNA sequences of genes fairly easily with the PCR (polymerase chain reaction) and direct sequencing methods. Using these techniques, many data concerning variation between and within species have been obtained. These data stimulated developments of the genealogy theory in population genetics resulting in various neutrality tests based on DNA data (HUDSON, 1990). In the present paper, I review these developments first and explain what discrepancies to the predictions of the neutral theory are now observed. Then, as one candidate model for

explaining the DNA data, multicodon nearly neutral mutation models are stated and recent analyses on them will be explained..

## 2 DNA data and the neutrality tests

First, I briefly explain the organization of genomes, the genetic materials of organisms. A genome is composed of sequences of DNA. For our purposes, it is enough to recognize them as sequences of characters called bases, A, G, T and C, corresponding to different nucleotides (for more details, see texts of molecular genetics). For example, the human genome consists of about $3 \times 10^9$ bases and the genome of a bacteria, *E. coli*, is about $5 \times 10^6$ bases. Only a portion of the genome codes for proteins or RNAs and other parts are called non-coding regions. Usually a protein is encoded by stretches of DNA called exons. They are interrupted by non-coding regions called introns. Three consecutive bases code for one amino acid and this unit is called a codon. Since there are 20 kinds of amino acids used in organisms and a codon can specify $4^3 = 64$ kinds, some base changes do not result in changes of amino acids. Such changes are called silent changes. Those changes causing amino acid changes are called replacement changes.

As far as population genetics is concerned, the improvements brought up by the DNA technologies are two-fold. First, by knowing the sequence, we can classify changes between and within populations as either silent (coding or non-coding) or replacement. Second, multiple changes in a gene can be identified. Such information was not at hand in the protein polymorphism era of 60's and 70's in population genetics. Changes at multiple sites in a gene enable us to infer the genealogical structure of genes in populations (see Fig. 1) and promoted the development of the genealogy theory (see reviews TAVARE, 1984; HUDSON, 1990). In this theory, descents of multiple genes are followed in the direction to the past

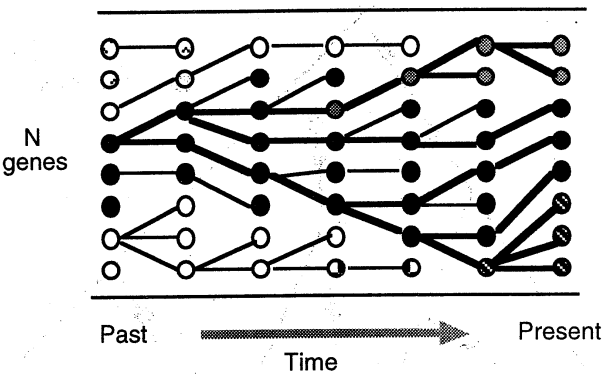



Figure 1: Gene genealogy in a random mating population. Circles represent genes.

and the genealogical structure (i.e., topology and times to coalescences) is probabilistically characterized. If we just look at gene frequencies in the present generation, the data types in Fig. 1 is $(2/8, 3/8, 3/8)$. This corresponds to to the data type in the protein polymorphism era utilizing the electrophoresis. However, if we can sequence these genes and know the differences among genes at multiple sites, the shape of the genealogy of the sampled genes can be estimated. Bellow I explain several neutrality tests developed to utilize such information.

The first test is that based on the variance of numbers of substitutions. Assume that sequences of a gene from multiples species (one from each species) are known. By comparing them, we can estimate the numbers of substitutions on the branches of the genealogical tree. For simplicity, consider $n$ species that diverged at the same time in the past. Let $X$ be the number of substitutions that occurred in one lineage after the diversification of the species. In this case, we can estimate the average, $\mathrm{E}[X]$, and variance, $\mathrm{Var}[X]$, of the number of substitutions in the gene after the diversification of the species. Under the neutrality assumption, the number of substitutions on each branch is expected to have approximately a Poisson distribution. Thus, the ratio of the variance to the average, called the dispersion index, $I = \mathrm{Var}[X]/\mathrm{E}[X]$, is expected to be one (For a general account of the dispersion index in molecular evolution, see TAKAHATA, 1987).

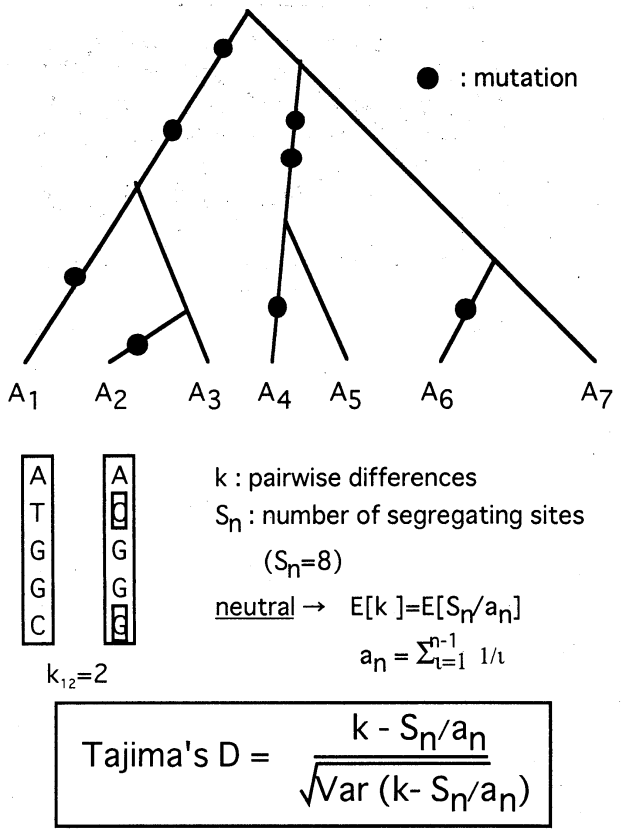



Figure 2: Tajima's test statistics. The genealogy of sampled sequences $A_1, \ldots, A_7$ is shown.

GILLESPIE (1989) and OHTA (1995) estimated the dispersion index of replacement substitutions using 20 and 49 mammalian genes and they obtained estimates of 6.95 and 5.6, respectively, significantly larger than one. They both concluded from these data that some selection is at work for replacement substitutions in mammals. Recently, ZENG et al. (1998) estimated $I$ using 24 genes of fruit flies, Drosophila, and obtained 1.6 as an estimate of $I$ for replacement substitutions. At present estimates of dispersion indices for other species are not available.

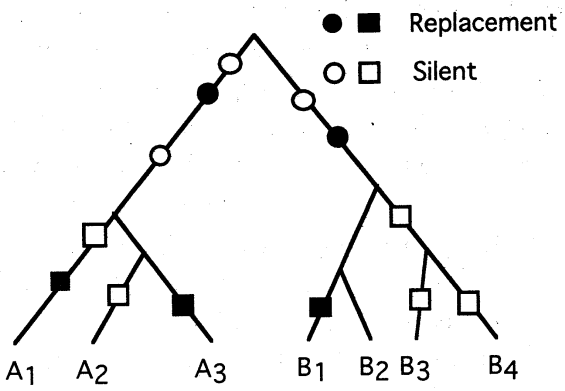
The second test is based on population polymorphism data. Assume that there are infinitely many sites and there is no recombination among them (the infinite site model without recombination of WATTERSON 1975). Consider that we sampled $m$ sequences from a population (see Fig. 2). If a site is variable, that is, there are variant nucleotides in some sequences at the site, the site is called a segregating site. Let $S_m$ be the number of segregating sites in $m$ sampled sequences. In the infinite site model without recombination, $S_m$ is the number of mutations in the whole genealogy (see Fig. 2). In a random mating equilibrium population with size $N$, if mutations are all neutral with respect to selection and mutation rate is $u$, the expected number, $\mathrm{E}[S_m]$, of segregating sites is expressed as

$$\mathrm{E}[S_m] = a_m \theta,$$

where $a_m = \sum_{i=1}^{m-1} 1/i$ and $\theta = 4Nu$ (WATTERSON, 1975). Next let $k_{ij}$ be the number of different sites between sequences $i$ and $j$. Its expected value is $\mathrm{E}[S_2]$ and thus $\mathrm{E}[k_{ij}] = \theta$. TAJIMA (1989a) introduced a neutrality test noting that $S_m/a_m$ and $k = (1/m(m-1)) \sum_i \sum_{j \neq i} k_{ij}$ both estimate $\theta = 4Nu$. His statistics called Tajima's $D$ is defined as

$$D = \frac{k - S_m/a_m}{\sqrt{u_T S_m + v_T S_m^2}},$$

where the denominator of the right-hand side of the equation is an estimator of the variance of the numerator (see TAJIMA, 1989a, for the expressions of $u_T, v_T$). This statistics was shown to be approximately distributed as Beta with mean zero and variance one (TAJIMA, 1989a) and he proposed to use this statistics to test the neutrality (see SIMONSEN et al., 1996, for distributional properties and power of the test). If there are more rare variants than expected under the neutrality, $k$ is not much affected but $S_m$ is expected to increase. Thus, the numerator is likely to be minus. On the other hand, if there are more high-frequency variants, $k$ increases and $S_m$ decreases resulting in plus $D$. The former situation occurs if mutations are deleterious or there was a recent bottleneck of population size. The latter situation is expected if there is balancing selection keeping the variant frequencies at high levels or isolation of populations (see TAJIMA, 1989b,c). TAJIMA's test was applied to many DNA polymorphism data of Drosophila. At two out of 38




| 2 x 2 table | fixed | polymorphic |
|---|---|---|
| replacement | 2 ( ● ) | 3 ( ■ ) |
| silent | 3 ( ○ ) | 5 ( □ ) |

Figure 3: MK test. The genealogy of sampled sequences $A_1, \ldots, A_3$ from species A and $B_1, \ldots, B_4$ form species B is shown with mutations.

loci, $D$ was significantly minus at the 5% level. Another neutrality test based on a similar idea is that of FU AND LI (1993). They utilize the number, $\eta_s$, of singleton sites and defined two statistics,

$$D^* = \frac{S_a/a_m - \eta_s(\frac{n-1}{n})}{\sqrt{u_{D*}S_a + v_{D*}S^2}}$$

$$F^* = \frac{k - \eta_s(\frac{n-1}{n})}{\sqrt{u_{F*}S + v_{F*}S^2}}.$$

The expectation of the numerator is zero because $\mathrm{E}[\eta_s] = n\theta/(n-1)$. The expressions for $u_{D*}, v_{D*}, u_{F*}, v_{F*}$ are listed in SIMONSEN et al. (1996).

The third type of tests utilizes DNA variation between and within populations. Here, I explain the MK test proposed by MCDONALD AND KREITMAN (1991). In this test, multiple sequences are sampled from more than one species (see Fig. 3). In the figure, three and four sequences are sampled from species A and B, respectively. After an alignment of the sequences, sites are classified by two criteria. For simplicity, we assume that there is no recombination among sites and at most only one change occurs at each site in the whole genealogy. The latter condition is mostly satisfied unless the two species are not distantly related. First, a site is called fixed if there are no variants within species but different nucleotides are fixed in the two species. In terms of the genealogy, mutations having occurred in the branch connecting the genealogies of the two species (circles in the figure) are those at fixed sites. If the site has a variant in either species, it is called polymorphic (squares in the figure). Second, a site is called replacement if the change causes a change of amino acid. Otherwise, it is called silent. All segregating sites are classified by these two criteria and we obtain a $2 \times 2$ table (see the lower part of the figure). Under the neutrality, the ratios of replacement substitutions to silent substitutions at fixed and polymorphic sites are expected to be the same. We can test this by applying a goodness of fit test to the table. In the MK test, replacement and silent sites are assumed to have the same genealogy as depicted in Fig. 3 since there is no recombination. If two types of sites recombine freely, we need to take into account the difference of genealogies. In another test called the HKA test (HUDSON, KREITMAN AND AGUADE 1987), variations within and between two populations (species) at two independent loci, within which there is no recombination, are compared. Because genealogies of the two independent loci differ, we need to take into account this stochastic factor for the goodness of fit test as was done by HUDSON, KREITMAN AND AGUADE (1987). These two tests of the neutrality were applied to Drosophila nuclear DNA data and it was found that about one-half of the loci examined did not conform to the expectation of the neutral hypothesis by either one of the test (MORIYAMA AND POWELL 1996). In addition, the MK test was applied to the mitochondrial DNA data in Drosophila (BALLARD AND KREITMAN, 1994; RAND et al., 1994), mouse (NACHMAN et al., 1994) and human (NACHMAN et al., 1996). In all cases, the neutrality was rejected. Furthermore, in these cases, excess replacement substitutions are found at polymorphic sites.

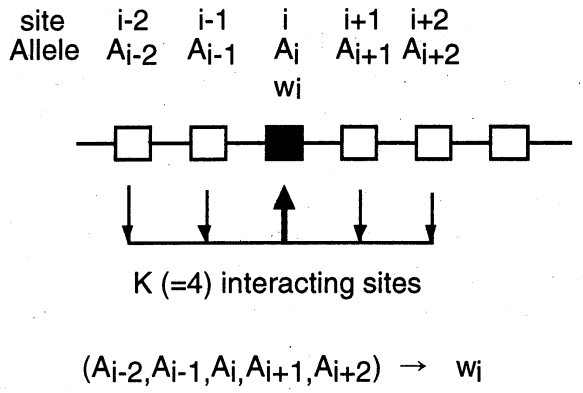



Figure 4: KAUFFMAN's NK model. The number of interacting amino acids is $K = 4$.

Another feature of DNA variation recently found is a reduction of variation (site heterozygosity or nucleotide diversity) at silent sites within population in regions of low recombination in Drosophila (BEGUN AND AQUADRO, 1992). Since variation between populations (species) is not reduced in regions of low recombination, we can not explain this pattern by postulating low mutation rate in regions of low recombination.

Although the neutrality was rejected at not all loci, there are many loci where DNA variation pattern is not compatible with the neutral expectation. Therefore, we need to explore other possibilities than the strict neutral model to explain evolution at the molecular level. In the next section, I describe one such effort introducing weak selection (the nearly neutral mutation model, see OHTA, 1972, 1973, 1992).

# 3 Nearly neutral mutation model

Here, we concentrate on evolution of proteins. A protein comprises of a sequence of amino acids. In the modeling of a protein, how mutation occurs and how fitness is determined should be specified. KAUFFMAN (1993) provides a general framework called the NK model to study evolution of proteins. In the NK model, a protein is assumed to have $N$ amino acids. The gene coding for the protein consists of $N$ sites, each site being able to specify one out of $L$ amino acids (see Fig. 4). $K$ represents the number of interacting amino acid sites and this will be explained later. At each site, an allele (amino acid) can mutate to any other alleles with equal probabilities. The allelic state of the gene is determined by what amino acids occupy respective amino acid sites in the gene. Fitness of an allele of the gene is determined as follows: Fitness, $w_i$, of an amino acid site $i$ is determined by what amino acids occupy the site and $K$ other sites in the NK model. In Fig. 4, the combination of the amino acid at the site ($A_i$) and those at $K = 4$ neighboring sites ($A_{i-2}, A_{i-1}, A_{i+1}, A_{i+2}$) determine the fitness of the site $w_i$. The fitness of the combination of amino acids at each site $i$ is determined by independently drawing numbers from a specified distribution, $f(s)$, at the start and its stays constant through time. The fitness, $w$, of the gene is the average of $w_i$s,

$$w = \frac{1}{N}\sum_{i=1}^{N} w_i.$$

If there is no interaction ($K = 0$), the fitness landscape has one peak and a population moves toward that peak. As $K$ increases, the fitness landscape becomes more rugged (KAUFFMAN, 1993).

Although this way of introducing interaction among amino acid sites is just one way of doing so and the original intention of introducing the NK model was more toward understanding complexity, the NK model provides a starting point for studying molecular evolution of protein. Past nearly neutral mutation models are special cases of the NK model. OHTA (1977) and KIMURA (1979) studied a protein model in which each mutation causes a shift of fitness with some specified distribution (the shift model). This corresponds to the $N = \infty, K = 0$ NK model. In this model, the average fitness increases or decreases indefinitely depending on whether there is positive mass in the specified distribution. On the other hand, OHTA AND TACHIDA (1990) and

TACHIDA (1991) studied a model in which the fitness of the gene is determined from a specified distribution when mutation occurs (the fixed model or house-of-cards model of KINGMAN, 1978). If there are $N$ sites each with $L$ alleles, there are $N \times L$ allelic states accessible from any one state by one mutation. With $K = N - 1$ in the NK model, fitnesses of those accessible states are randomly assigned at the outset. As $N$ goes to infinity, the number of accessible states becomes infinite and the model is expected to converge to the house-of-cards model.

Several studies examined the behavior of these nearly neutral mutation models paying attention to the statistics mentioned in the previous section. In the house-of-cards model, $\alpha = 2N\sigma$ (hereafter, we use $N$ for designating population size and $\sigma$ for the standard deviation of the mutational effect on fitness) determines most of the model behavior (TACHIDA, 1991, 1996). IWASA (1993), and GILLESPIE (1994a) found that the dispersion index becomes very large for $\alpha > 1$ although the realistic range might be small because substitutions almost cease if $\alpha > 4$ in the house-of-cards model. If size of population changes, the dispersion index becomes large and still some substitutions occur in the house-of-cards model (ARAKI AND TACHIDA, 1997). The average of TAJIMA's $D$ is negative both in the shift and house-of-cards model (GILLESPIE, 1994b 1997). OHTA (1997, 1998) investigated the NK model with $K = 0, 2, 4$ and showed that the dispersion index is close to one and TAJIMA's $D$ is minus. However, the distributional properties of TAJIMA's $D$ and the MK test have not been investigated. Because large amount of data are now accumulating, it is necessary to characterize behavior of the models comprehensively in terms of these statistics. As a start, I chose the simplest nearly neutral mutation model mimicking a protein coding-gene structure, i. e., the NK model with $K = 0$ (the multi-codon model) and investigated the pattern of variation with regard to these statistics for the neutrality tests (TACHIDA, 1999).

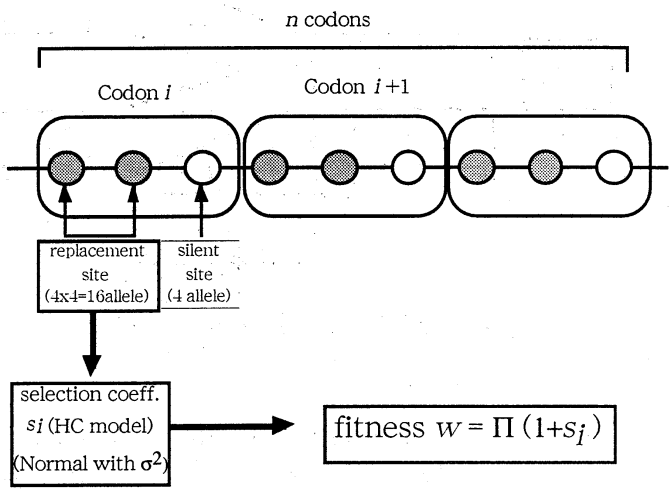



Figure 5: Multicodon model.

# 4 Multicodon model

Consider a gene consisting of $3n$ nucleotide sites each with four alleles (Fig. 5). No recombination is assumed among nucleotide sites. Each of three consecutive sites is called a codon. For simplicity the first two sites of a codon is assumed to specify an amino acid and the third site is a silent site. Because there are four alleles (A, T, G, C) per site, a codon specifies one from sixteen amino acids. In real organisms, $4^3 = 64$ codons specify 20 amino acids and a stop signal with redundancy mostly at the third site. Mutation occurs $u$ per site per generation and the probabilities of changing to other three alleles are equal. For fitness, independence among sites is assumed. For $i$th codon, the selection coefficient $s_i$ is assigned from a specified distribution $f(s)$ (a normal distribution with a mean zero and variance $\sigma^2$ in the following) to each of 16 amino acids and they stay constant through time. The fitness of a gene $w$ is defined multiplicatively as

$$w = \prod_{i=1}^{n}(1 + s_i).$$

This model differs from the NK model with $N = 3n, K = 1, L = 4$ in the following points. First, interaction between sites is only for the first and second sites in a codon and there is no interaction among codons. In the NK model

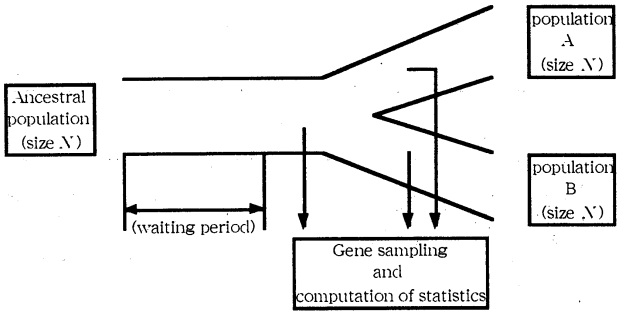



Figure 6: Population model.

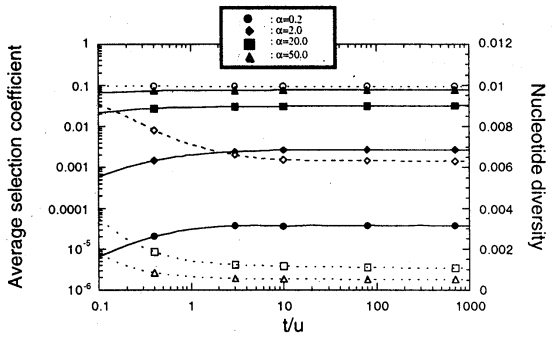



Figure 7: Average selection coefficients and nucleotide diversity in one-codon model. Solid lines and dotted lines represent average selection coefficients and nucleotide diversity, respectively.

with $K = 1$, if interacting sites are taken to be one and the next one, all sites consequently interact. Second, there are silent sites. As described before, some neutrality tests uses information from both silent and replacement variation. Thirdly, the fitness is defined multiplicatively while it is defined additively in the NK model. Since relative fitnesses determine the dynamics of gene frequencies, the effect of changing one amino acid on fitness is independent on the average fitness in the multiplicative fitness models.

The population starts from $N$ genes with the same allelic state (see Fig. 6). The Wright-Fisher model (see, for example, EWENS, 1979; CROW AND KIMURA, 1970) is assumed with discrete generations and constant size $N$ throughout time. The initial allelic state was chosen by randomly assigning amino acids at respective codons. After a weighting period for stationarity (explained later), a gene is sampled from the population. The sequence of this gene is later used as an outgroup sequence to estimate numbers of substitutions. $10N$ generations after the first sampling, the population is split into two, A and B, with the same size $N$ and random mating continues within each population. At every $0.05/u$ generations after the split, $m$ genes are sampled from each population and their sequences are recorded up to $0.5/u$ generations. This range of time covers most of the data analyses conducted thus far.

This model is analyzed using computer simulation. However, before conducting simulation of the full multicodon model, a single-codon model was approximately analyzed to determine the weighting period and examine to what extent the single-codon model approximates the multicodon model. In the single-codon model, a gene consists of a single codon. Because the product of population size and mutation rate is small compared to one, we use the weak mutation approximation where the population is assumed to be monomorphic and represented by the fixed allele. Evolution is described by a 16-states Markov chain (see, for example, ZENG, TACHIDA AND COCKERHAM, 1989). With this approximation, it can be shown using KIMURA (1962)'s formula for fixation probabilities that the equilibrium distribution, $q_i$, for the population to be fixed with the $i$th allele is expressed as

$$q_i = \frac{\exp(2Ns_i)}{\sum_{j=1}^{16} \exp(2Ns_j)}, \qquad (1)$$

where $s_i$ is the selection coefficient of the $i$th allele (TACHIDA, unpublished results). Time dependent behavior of the average fitness, site heterozygosity (hereafter called nucleotide diversity) and number of substitutions were examined using computer simulation of the approximate Markov chain (for the method, see TACHIDA, 1996). By this approximation, computing time is reduced very much and we can obtain the result up to $1000/u$ generations with

large number of replications.

The result is shown in Fig. 7. From the figure, the average selection coefficient and nucleotide diversity seem to achieve stationary values by $5/u - 10/u$ generations. Although the values achieved are close to equilibrium values for $\alpha < 10$, this is not true for larger $\alpha$. For example, with $\alpha = 20$, the average selection coefficient at equilibrium computed from (1) is 0.0353 while it is 0.0301 at $t = 5/u$ and 0.0320 even at $t = 1000/u$. However, mutation rate per nucleotide site is about $10^{-8}$ or less per year. Thus, the approximate stationarity achieved in $5/u - 10/u$ generations is what we need to consider in protein evolution. From this consideration, the full multicodon model simulation was conducted with a weighting period of $5/u$ generations.

The multicodon simulation was carried out closely following the Wright-Fisher model with multinomial sampling done by the rejection method (see PRESS et al, 1988). First, we examined the average selection coefficient, $\bar{w}$, and nucleotide diversities at replacement site ($\pi_r$) and silent sites ($\pi_s$) and results are shown in Table 1. For comparison, the results of one-codon simulations are also shown. The agreement between the multicodon and single-codon simulations are very good for $\bar{w}$ and $\pi_r$. This justifies the use of the single-codon approximation for computing these quantities. Furthermore, this suggests that codons are approximately evolving independently. This may allow us to analyze the statistics described below with analytical approximations in future studies. The nucleotide diversity, $\pi_s$, at linked silent sites shows some interesting pattern. As $\alpha$ increases, $\pi_s$ first decreases and then increases. The maximum reduction from the neutral ($\alpha = 0$) value is found for $\alpha = 5.0$ and it is 15% with $u = 10^{-5}$. The relative reduction increases as $u$ becomes large. This suggests that intermediate intensity of selection can be one explanation for the reduction of the nucleotide diversity at linked silent sites mentioned before.

Next we investigated the number of substitutions. To estimate numbers of substitutions, we sampled a gene $10N$ generations before the split of the population and this is used as an outgroup. After the split, one gene each is sampled from populations A and B (see Fig. 6). The numbers of substitutions between all pairs of the three genes are estimated by the JUKES-CANTOR method (JUKES AND CANTOR, 1968) based on their differences. From these numbers of substitutions, we can draw a gene genealogy and estimate the numbers of substitutions on the branches leading to the gene from A and gene from B. Let $X_A$ and $X_B$ be these estimated numbers of substitutions. This procedure is the one usually taken in estimations of substitution rates. We can estimate the mean and variance of numbers of substitutions by averaging $(X_A + X_B)/2$ and $(X_A - X_B)^2/2$ over replications, respectively. As shown by BULMER (1989), this way of estimating the variance introduces bias when the number of substitutions per site becomes large. Thus, in the estimation, we used his correction factor. Some of the results are shown in Table 2. The average number of substitutions decreases as selection becomes stronger. Al-

Table 2: The average number of substitutions and dispersion index.

| | $t^a = 0.1/u$ | | $t = 0.5/u$ | |
|---|---|---|---|---|
| $\alpha$ | sub.[b] | $I^c$ | sub.[b] | $I^c$ |
| 0.0 | 0.1055 | 1.023 | 0.5120 | 0.989 |
| 0.2 | 0.1036 | 0.970 | 0.4985 | 1.046 |
| 2.0 | 0.0558 | 1.013 | 0.2160 | 1.137 |
| 10.0 | 0.0093 | 0.917 | 0.0303 | 1.022 |
| 20.0 | 0.0044 | 1.055 | 0.0142 | 1.141 |
| 50.0 | 0.0029 | 1.073 | 0.0070 | 1.344 |

[a] Time after the split of the population.
[b] Average number of substitutions.
[c] Dispersion index.
The values are computed from the outputs of 1000 replications with population size $N = 500, u = 10^{-5}, n = 100$.

Table 1: The average fitness and nucleotide diversity at $5.6/u$ generation

| $\alpha$ | $u(\times 10^5)$ | $\bar{w}^a$ | $\bar{w}$ (1 codon)$^b$ | $\pi_r^a$ | $\pi_r$ (1 codon)$^b$ | $\pi_s^a$ |
|---|---|---|---|---|---|---|
| 0.0 | 1 | 1.0000 | 1.0000 | 0.00979 | 0.01000 | 0.01007 |
| 0.2 | 1 | 1.0036 | 1.0038 | 0.00989 | 0.00993 | 0.00970 |
| 2.0 | 1 | 1.2582 | 1.2974 | 0.00652 | 0.00645 | 0.00907 |
| | 4 | 1.2107 | 1.2974 | 0.02361 | 0.02581 | 0.02971 |
| 5.0 | 1 | 2.1021 | 2.1325 | 0.00353 | 0.00350 | 0.00857 |
| 20.0 | 1 | 20.2720 | 19.9891 | 0.00114 | 0.00119 | 0.00889 |
| | 4 | 20.9628 | 19.9891 | 0.00454 | 0.00476 | 0.02823 |
| 50.0 | 1 | 1446.5668 | 1455.6000 | 0.00054 | 0.00056 | 0.00931 |

[a] Computed from the multicodon simulations with 1000 replications. $N = 500, n = 100$.
[b] Computed from the single-codon simulations with $10^5$ replications.

though the substitutions almost stop for $\alpha > 4$ in the house-of-cards model with infinite number of alleles, the average number of substitutions for $\alpha = 10$ is about one-tenth of the neutral value in the present model. This is due to the finiteness of the number of alleles in one codon. The maximum selection coefficient exists in the finite allele models while it does not in the infinite allele model if mutant effects on fitness is without bound like those from a normal distribution. The dispersion index, $I = \mathrm{Var}[X]/\mathrm{E}[X]$, was estimated from the average and variance of the number of substitutions. In the period examined, the dispersion index is generally close to one. The reason for low dispersion indices in the present model is considered to be due to two factors. First, the number of substitutions is averaged over codons. In codons where differences of selection coefficients are small, the number of substitutions is large and thus its contribution is larger than those from other codons. In these codons, substitutions occur more like those in the neutral case and would be regular. Second, it takes time for the dispersion index to be larger in the house-of-cards model (TACHIDA, 1996). Here, the time is short in terms of $1/u$ generations. This pattern in dispersion indices is different from that found in mammals.

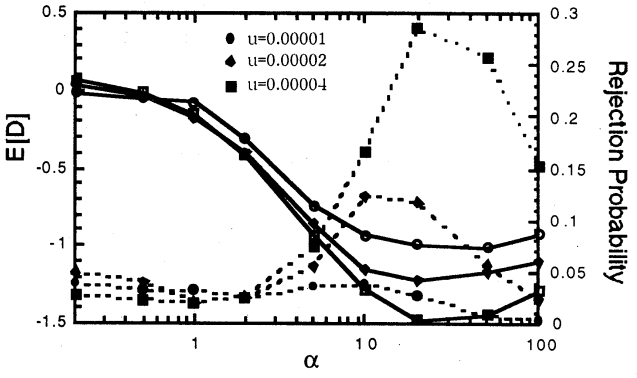



Figure 8: Tajima's $D$ at replacement sites. Averages and rejection probabilities ($P = 0.05$) are shown by solid and broken lines, respectively. Parameters are $N = 500, n = 100, m = 50$.

Next we examined TAJIMA's $D$. It is estimated by first sampling multiple genes from a population and then estimating the average number of differences between genes, $k$, and the number of segregating sites, $S_m$. The average $D$ and the probability of rejection of the neutrality at the 5% level were estimated. Some of the results at replacement sites are shown in Fig. 8. As shown in the figure, E[$D$] starts to decreases from zero as $\alpha$ becomes larger than one and converges to around minus one. Although E[$D$] becomes minus, the rejection probability is low and takes the maximum value around 0.3 when $u = 0.000004, \alpha = 20$.

The probability becomes smaller as mutation rate decreases and for $u = 0.00001$ it never becomes larger than 0.05. As selection becomes very strong ($\alpha = 100$), $|\mathrm{E}[D]|$ decreases and the probability drops. This is because the population becomes monomorphic and not much variation is left reducing the detection rate of selection. Sample size also affects the rejection probability. For example, the rejection probabilities for $\alpha = 20$ are 0.122, 0.286 and 0.402 with $m = 20, 50, 100$, respectively. If the number, $n$, of sites increases, the rejection probability increases. Approximately $nu$ determines the rejection probability. This is expected because each codon evolves independently and the nucleotide diversity is very low. At silent sites $|\mathrm{E}[D]|$ is close to zero and the rejection probability is low. FU AND LI's $D^*$ and $F^*$ behave similarly (data not shown). These statistics based on DNA polymorphism data generally take negative values under the present model and the rejection probability is low unless $u$ or $n$ is large.

Finally, we carried out the MK test sampling multiple genes from two populations. Results are shown in Table 3. The rejection probability of the neutrality is shown changing $\alpha, u, n, m, t$. Time is measured defining the split time of the two populations as zero. In the MK test, there are two ways the null hypothesis that the replacement/silent ratio at the polymorphic sites are the same as that at fixed sites is rejected. Define the excess replacement changes at polymorphic sites as "rejA" and the converse is defined as "rejB". Let $P_{\mathrm{rejA}}$ and $P_{\mathrm{rejB}}$ be the respective probabilities of these two events. The table presents these two probabilities when the null hypothesis is rejected at the 5% level. First of all, note that the rejection is almost always "rejA". This is a characteristic outcomes of the present nearly neutral mutation model with constant population size. Such pattern is rarely found in nuclear genes (MORIYAMA AND POWELL, 1996, but see also MIYASHITA et al., 1998) but found many times in mitochondrial DNA (BALLARD AND KREITMAN, 1994; RAND et al., 1994; NACHMAN et al, 1994 and 1996; HASEGAWA et al., 1998). The MK test is sensitive in detecting weak selection. Even for $\alpha = 2.0$, the rejection probability is 0.1 when sample size is $m = 50$. Also the rejection probability is significant even with small sample size for $\alpha = 10$. However, as selection becomes stronger ($\alpha = 50$), the rejection probability becomes smaller. This is in sharp contrast to TAJIMA's test where the rejection probability is fairly high for $\alpha = 50$ but very low for $\alpha = 2$. The reason for the low rejection probability of the MK test when selection is strong (large $\alpha$) is the rapid decrease of the fixed replacement sites. The power of the MK test depends on the number of replacement substitutions at fixed sites. Thus, the time after the split also affects the rejection probability. The rule that $nu$ determines the rejection probability found in TAJIMA's test does not hold in the MK test. The rejection probability increases as $n$ increases and with $n = 400, u = 0.00001$, the selection is detected very efficiently.

# 5 Conclusion

Recent accumulation of data of DNA variation and advances in analyzing these data were reviewed. These new data suggest some inadequacies of the strict neutral model of molecular evolution, especially of protein evolution, urging us to study other models, possibly involving selection (GILLESPIE, 1993, 1994b). As one such effort for understanding protein evolution, a simple multicodon nearly neutral mutation model was proposed and its behavior was investigated with special attention to the statistics for testing the neutrality. With the assumption of constant population size and no recombination within a gene, the present model predicts the outcome of the neutrality tests as follows:

1. The dispersion index is close to one.

2. TAJIMA's and FU AND LI statistics have negative values but rejection of the neu-

Table 3: Results of the MK test with various sample size and divergence time.

| | | | | $t = 0.1/u$ | | $t = 0.15/u$ | |
|---|---|---|---|---|---|---|---|
| $n$ | $u^a$ | $\alpha$ | $m^b$ | $P^c_{\text{rejA}}$ | $P^d_{\text{rejB}}$ | $P^c_{\text{rejA}}$ | $P^d_{\text{rejB}}$ |
| 100 | 1 | 2 | 10 | 0.067 | 0.008 | 0.067 | 0.003 |
| | | | 20 | 0.088 | 0.007 | 0.094 | 0.002 |
| | | | 50 | 0.116 | 0.004 | 0.100 | 0.001 |
| | | 10 | 10 | 0.219 | 0.000 | 0.271 | 0.000 |
| | | | 20 | 0.334 | 0.000 | 0.405 | 0.000 |
| | | | 50 | 0.467 | 0.000 | 0.602 | 0.000 |
| | | 50 | 10 | 0.042 | 0.000 | 0.073 | 0.000 |
| | | | 20 | 0.066 | 0.000 | 0.139 | 0.000 |
| | | | 50 | 0.146 | 0.000 | 0.264 | 0.000 |
| | 4 | 2 | 50 | 0.073 | 0.005 | 0.090 | 0.003 |
| | | 10 | 50 | 0.530 | 0.000 | 0.684 | 0.000 |
| 400 | 1 | 2 | 50 | 0.134 | 0.000 | 0.161 | 0.000 |
| | | 10 | 50 | 0.901 | 0.000 | 0.940 | 0.000 |

[a] $u$ times $10^5$. [b] Sample size.
[c] Probability of rejection at the 5% level with excess replacement at polymorphic sites.
[d] Probability of rejection at the 5% level with excess replacement at fixed sites.
The values obtained from 1000 replications with $N = 500$.

trality occurs when $\alpha$ is more than ten and mutation rate or the number of codons are large.

3. The MK test can detect selection of this type if $\alpha$ is more than two and the direction of the rejection is always that of excess replacement polymorphisms.

4. A reduction of nucleotide diversity at linked silent sites results under this model but the maximum reduction occurs at intermediate strength of selection (around $\alpha = 5$).

As mentioned before, mammalian nuclear genes have large dispersion indices (GILLESPIE, 1989; OHTA, 1995) and the direction of the rejection in the MK test of Drosophila nuclear genes is always in the other direction, i.e., excess replacement fixed sites. Thus, the present model can not explain these nuclear gene data. However, if we introduce changes of population size, the dispersion index is expected to increase (ARAKI AND TACHIDA, 1997) and excess replacement fixations may occur in the MK test comparison. Quantifying the effects of changing population size is one immediate problem for future researches. Furthermore, silent sites are not necessarily neutral as reported in Drosophila (AKASHI, 1995) and background selection (CHARLESWORTH ET AL., 1993, 1995; HUDSON AND KAPLAN, 1995; NORDBORG et al., 1996) and hitchhiking (KAPLAN AND HUDSON, 1989; BRAVERMAN et al., 1995) may affect variation in the protein coding genes. Extensions incorporating these factors and its mathematical analysis (like that by SAWYER AND HARTL, 1992) are one path to enhance our understanding of the protein evolution.

# References


AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. Genetics 139: 1067-1076.

ARAKI, H., AND H. TACHIDA, 1997 Bottleneck effect on evolutionary rate in the nearly neutral mutation model. Genetics 147: 907-914.

BALLARD, J. W. O., AND M. KREITMAN, 1994 Unraveling selection in the mitochondrial genome of Drosophila. Genetics 138: 757-772.

BEGUN, D. J., AND C. F. AQUADRO, Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster.* nature 356: 519-520.

BRAVERMAN, J. M., R. R. HUDSON, N. L. KAPLAN, C. H. LANGLEY, AND W. STEPHAN, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140: 783-796.

BULMER, M., 1989 Estimating the variability of substitution rates. Genetics 123: 615-619.

CHARLESWORTH, B., M. T. MORGAN AND D. CHARLESWORTH, 1993 The effects of deleterious mutations on neutral molecular variation. Genetics 134: 1289-1303.

CHARLESWORTH, D., B. CHARLESWORTH AND M. T. MORGAN, 1995 The pattern of neutral molecular variation under background selection model. Genetics 141: 1619-1632.

CROW, J. F. AND M. KIMURA, 1970 *An Introduction to Population Genetic Theory.* Harper and Row, New York.

DYKHUIZEN, D. E. AND D. L. HARTL, 1980 Selective neutrality of 6PGD allozymes in *Escherichia coli.* Genetics 96: 801-817.

EWENS, W. J., 1972 The sampling theory fro selectively neutral alleles. Theor. Popul. Biol. 3: 87-112.

EWENS, W. J., 1979 *Mathematical Population Genetics.* Springer, New York.

FU, Y.-X., AND W.-H. LI, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693-709.

GILLESPIE, J. H., 1977 Sampling theory for alleles in a random environment. Nature 266: 443-445.

GILLESPIE, J. H., 1989 Lineage effects and the index of dispersion of molecular evolution. Mol. Biol. Evol. 6: 636-647.

GILLESPIE, J. H., 1991 *The Causes of Molecular Evolution.* Oxford University Press, New York.

GILLESPIE, J. H., 1993 Substitution processes in molecular evolution. I. Uniform and clustered substitutions in a haploid model. Genetics 134: 971-981.

GILLESPIE, J. H., 1994a Substitution processes in molecular evolution. III. deleterious alleles. Genetics 138: 943-952.

GILLESPIE, J. H., 1994b Alternatives to the neutral theory, pp. 1-17 in *Non-Neutral Evolution,* edited by B. Golding. Chapman & Hall, New York.

GILLESPIE, J. H., 1997 Junk ain't junk does: neutral alleles in a selected context. Gene 205: 291-299.

HASEGAWA, M., Y. CAO AND Z. YANG, 1998 Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. Mol. Biol. Evol. 15: 1499-1505.

HUDSON, R. R., 1990 Gene genealogies and coalescent process. Oxford Surveys Evol. Biol. 7: 1-44.

HUDSON, R., R., KREITMAN, M., AND M. AGUADE, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

HUDSON, R. R. AND N. L. KAPLAN, 1995 Deleterious background selection with recombination. Genetics 141: 1605-1617.

IWASA, Y., 1993 Overdispersed molecular evolution in constant environments. J. Theor. Biol. 164: 373-393.

JUKES, T. H., AND C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21-132 in *Mammalian Protein Metabolism,* edited by H. N. MURO. Academic Press, New York.

KAPLAN, N. L., R. R. HUDSON, AND C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics 123: 87-899.

KAUFFMAN, S. A., 1993 *The Origins of Order.* Oxford University Press.

KIMURA, K., 1962 On the probability of fixation of mutant genes in a population. Genetics 47: 713-719.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature 217: 624-626.

KIMURA, M., 1979 A model for effectively neutral mutations in which selective constraint is incorporated. Proc. Natl. Acad. Sci. USA 76: 3440-3444.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KINGMAN, J. F. C., 1978 A simple model for the balance between selection and mutation. J. Appl. Probab. 15: 1-12.

KREITMAN, M., 1996 The neutral theory is dead. Long live the neutral theory. BioEssays 18: 678-683.

LEWONTIN, R. C., 1974 *The Genetic Basis of Evolutionary Change.* Columbia University Press, New York.

MCDONALD, J. H., AND KREITMAN, M., 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila.* Nature 351: 652-654.

MIYASHITA, N. T., A. KAWABE, H. INNAN AND R. TERAUCHI, 1998 Intra- and interspecific DNA variation and codon bias of the alcohol dehydrogenase (*Adh*) locus in *Arabis* and *Arabidopsis* species. Mol. Biol. Evol. 15: 1420-1429.

MORIYAMA, E. N., AND J. R. POWELL, 1996 Intraspecific nuclear DNA variation in *Drosophila.* Mol. Biol. Evol. 13: 261-277.

MUKAI, T., M. ICHINOSE AND H. TACHIDA, 1980 Selection for viability at loci controlling protein polymorphisms in *Drosophila melanogaster* is very weak at most. Proc. Natl. Acad. Sci. USA 77: 4857-4860.

NACHMAN, M. W., S. N. BOYER, AND C. F. AQUADRO, 1994 Nonneutral evolution at the mitochondrial NADH dehydrogenase subunit 3 gene in mice. Proc. natl. Acad. Sci. USA 91: 6364-6368.

NACHMAN, M. W., W. M. BROWN, M. STONEKING AND C. F. AQUADRO, 1996 Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics 142: 953-963.

NORDBORG, M., B. CHARLESWORTH AND D. CHARLESWORTH, 1996 The effect of recombination on background selection. Genet. Res. Camb.67: 159-174.

OHTA, T., 1972 Population size and rate of evolution. J. Mol. Evol. 1: 305-314.

OHTA, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature 246: 96-98.

OHTA, T., 1977 Extension to the neutral mutation random drift hypothesis, pp. 148-167 in *Molecular Evolution and Polymorphism,* edited by M. Kimura. National Inst. Genet., Mishima.

OHTA, T., 1992 The nearly neutral theory of molecular evolution. Ann. Rev. Syst. Ecol. 23: 263-286.

OHTA, T., 1995 Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J. Mol. Evol. 40: 56-63.

OHTA, T., 1996 The current significance and standing of neutral and nearly neutral theories. BioEssays 18: 673-677.

OHTA, T., 1997 Role of random genetic drift in the evolution of interactive systems. J. Mol. Evol. 44: S9-S14.

OHTA, T.. 1998 Evolution by nearly-neutral mutations. Genetica 102/103: 83-90.

OHTA, T. AND H. TACHIDA, 1990 Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics 126: 219-229.

RAND, D. M., M. DORFSMAN, AND L. M. KAN, 1994 Neutral and nonneutral evolution of Drosophila mitochondrial DNA. Genetics 138: 741-756.

PRESS, W. H., B. P., FLANNERY, S. A. TUEKOLSKY, AND W. T. VETTERLING, 1988 *Numerical Recipes in C.* Cambridge University Press, Cambridge.

SAWYER, S. A., AND D. L. HARTL, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161-1176.

SIMONSEN, K. L., G. A. CHURCHILL AND C. F. AQUADRO, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics 141: 413-429.

TACHIDA, H., 1991 A study on a nearly neutral mutation model in finite populations. Genetics 128: 183-192.

TACHIDA, H., 1996 Effects of the shape of distribution of mutant effect in nearly neutral mutation models. J. Genet. 75: 33-48.

TACHIDA, H., 1999 Molecular evolution in multisite nearly neutral mutation model. (submitted).

TAJIMA, F., 1989a Statistical method for testing the neutral mutation hypothesis. Genetics 123: 585-595.

TAJIMA, F., 1989b DNA polymorphism in a subdivided population: the expected number of segregating sites in the two subpopulation model. Genetics 123: 229-240.

TAJIMA, F., 1989c The effect of change in population size on DNA polymorphism. Genetics 123: 597-601.

TAKAHATA, N., 1987 On the overdispersed molecular clock. Genetics 116: 169-179.

TAVARE, S., 1984 Line-of-descent and genealogical processes and their applications in population genetics. Theor. Popul. Biol. 25: 119-164.

WATTERSON, G. A., 1975: On the number of segregating sites in genetic models without recombination. Theor. Popul. Biol. 7: 256-276.

ZENG, L.-W., J. M. CAMERON, B. CHEN, AND M. KREITMAN, 1998 The molecular clock revisited: the rate of synonymous vs. replacement change in *Drosophila.* Genetica 102/103: 369-382.

ZENG, Z.-B., H. TACHIDA AND C. C. COCKERHAM, 1989 Effects of mutation on selection limits in finite populations with multiple alleles. Genetics 122: 977-984.